/

CS

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
09/918,952	07/31/2001	Philip Shi-Lung Yu	YOR9-2001-0363US1 (8728-5	2574

7590 09/25/2003

Frank Chau
F. CHAU & ASSOCIATES, LLP
Suite 501
1900 Hempstead Turnpike
East Meadow, NY 11554

EXAMINER	
FLEURANTIN, JEAN B	
ART UNIT	**PAPER NUMBER**
2172	

DATE MAILED: 09/25/2003

Please find below and/or attached an Office communication concerning this application or proceeding.

Office Action Summary	Application No.	Applicant(s)
	09/918,952	YU ET AL.
	Examiner	Art Unit
	Jean B Fleurantin	2172

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE _07_ MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133).
- Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on _31 July 2003_ .

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) _1-25_ is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) _1-6,9,16-21 and 23-25_ is/are rejected.

7)☒ Claim(s) _7,8,10-15 and 22_ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☒ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

11)☐ The proposed drawing correction filed on _____ is: a)☐ approved b)☐ disapproved by the Examiner.

If approved, corrected drawings are required in reply to this Office action.

12)☐ The oath or declaration is objected to by the Examiner.

Priority under 35 U.S.C. §§ 119 and 120

13)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____ .

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

14)☐ Acknowledgment is made of a claim for domestic priority under 35 U.S.C. § 119(e) (to a provisional application).

a) ☐ The translation of the foreign language provisional application has been received.

15)☐ Acknowledgment is made of a claim for domestic priority under 35 U.S.C. §§ 120 and/or 121.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☒ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO-1449) Paper No(s) _2_ .
4) ☐ Interview Summary (PTO-413) Paper No(s). _____ .
5) ☐ Notice of Informal Patent Application (PTO-152)
6) ☐ Other:

DETAILED ACTION

1. This Office Action is in response to the application filed on July 31, 2001, in which claims 1-25 are presented for examination.

Information Disclosure Statement

2. The information disclosure statement filed on July 31, 2001, complies with the provisions of MPEP 609. It has been placed in the application file. The information referred to therein has been considered as to the merits.

Drawings

3. The drawings filed on July 31, 2001 are approved by the Draftsperson under 37 CFR 1.84 or 1.152 as indicated in the "Notice of Draftperson's Patent Drawing Review," PTO-948.

Specification

4. In page 47 of the application, Applicant is advised to delete the underline title on the abstract portion, and the extraneous matter found at the bottom of the page. See MPEP 608.01(f).

Claim Rejections - 35 USC § 103

5. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

Claims 1- 6, 9, 16-21 and 23-25 are rejected under 35 U.S.C. 103(a) as being unpatentable over U.S. Patent No. 5,799,311 issued to Agrawal et al. (hereinafter "Agrawal") in view of Ramaswamy et al. (hereinafter "Rmaswamy") "Efficient Algorithms for Mining Outliers from Large Data Sets - 05/2000".

As per claim 1, Agrawal teaches a method for building a decision tree from an input data set, the input data set comprising records and associated attributes, the attributes including a class label attribute for indicating whether a given record is a member of a target class or a non-target class, the input data set being biased in favor of the records of the non-target class (see col. 3, lines 25-29 of Agrawal), the decision tree comprising a plurality of nodes that include a root node and leaf nodes (see col. 3, lines 40-41 of Agrawal). In particular, Agrawal discloses the claimed features of "constructing the decision tree from the input data set, including the step of partitioning each of the plurality of nodes of the decision tree, beginning with the root node, based upon multivariate subspace splitting criteria" (see col. 3, lines 40-43 of Agrawal, as a means for creating a decision tree is created by repeatedly splitting the records at each examined node starting with the root node, at any examined node a split test is determined to best separate the records at that node by record class and using the attribute lists, the node's records are split

according to the best split test into partitions of records to form child nodes of the examined node); and

"classifying and scoring the records, based upon the decision tree and the nearest neighbor set of nodes" (see col. 2, lines 44-46 of Agrawal, as a means for classifying the nodes into the high or low risk categories). Agrawal does not explicitly disclose the steps of computing distance functions for each of the leaf nodes; and identifying, with respect to the distance functions, a nearest neighbor set of nodes for each of the leaf nodes based upon a respective closeness of the nearest neighbor set of nodes to a target record of the target class. However, Agrawal discloses the use of "a nearest neighbor set of nodes for each of the leaf nodes based upon a respective closeness of the nearest neighbor set of nodes to a target record of the target class" (see col. 6, lines 52-54 of Agrawal, as records at each new leaf node are checked at block twenty three to see if they are of the same class). Ramaswamy, on the other hand, discloses an analogous system that teaches the claimed features "computing distance functions for each of the leaf nodes" as a means for using the square of the euclidean distance as the distance metric, (see Ramaswamy page 429, col. 2, lines 36-46), which is similar to the description provided by the applicant (specification on page 33, lines 10-20). Applicant should duly note, that Ramaswamy uses the distance of the $k^{th.}$ neighbor of the $n^{th.}$ outlier to define the neighborhood distance d (see Ramaswamy page 428, col. 1, lines 50-52). Therefore, it would have been obvious to a person of ordinary skill in the art at the time of the invention was made to combine Ramaswamy's distance function (page 429) with Agrawal's nearest neighbor set of node (col. 6, lines 52-54), in other to achieve computing distance functions for each of the leaf nodes; and identifying, with respect to the distance functions, a nearest neighbor set of nodes for each of the leaf nodes based upon a

respective closeness of the nearest neighbor set of nodes to a target record of the target class. One having ordinary skill in the art at the time the invention was made would have been motivated to utilize such combination as such would have allowed Agrawal's system the enhanced capability of improving the accuracy and the efficiency of the method for building space splitting decision tree, and provide performance of the partition based algorithm relatively unchanged, (see Ramaswamy page 437, col. 1, lines 17-18), therefore providing a quicker computation time of identifying small patterns for a given data analysis.

As per claim 2, Agrawal teaches, wherein said constructing step comprises the steps of forming a plurality of pre-sorted attribute lists, each of the plurality of pre-sorted attribute lists corresponding to one of the attributes other than the class label attribute (see col. 3, lines 34-37 of Agrawal, as the attribute lists for numeric which attributes are sorted based on attribute value and a decision tree is then generated by repeatedly partitioning the records according to record classes using the attribute lists); and

constructing the root node to including the plurality of pre-sorted attribute lists (see col. 3, lines 40-41 of Agrawal, as a decision tree which is created by repeatedly splitting the records at each examined node and starting with the root node).

As per claim 3, Agrawal teaches, wherein said forming step comprises the step of forming each of the plurality of pre-sorted attribute lists to include a plurality of entries, each of the plurality of entries comprising a record id for identifying a

record associated with the corresponding one of the attributes, a value of the corresponding one of the attributes, and a value of the class label attribute associated with the record (see col. 3, lines 30-34 of Agrawal, as a means for generating an attribute list for each attribute of the training records, each entry in the attribute list includes a value of that attribute, and the class label and record ID of the record from which the attribute value came from).

As per claim 4, Agrawal teaches, wherein said partitioning step partitions a current node from among the plurality of nodes of the decision tree, starting with the root node, until the current node includes only attributes that indicate membership in a same class (see col. 3, lines 36-39 of Agrawal, as a decision tree which is then generated by repeatedly partitioning the records according to record classes using the attribute lists and the final decision tree becomes the desired classifier in which the records associated with each leaf node are of the same class).

As per claim 5, Agrawal teaches, wherein said partitioning step partitions a current node from among the plurality of nodes of the decision tree, starting with the root node (see col. 3, lines 40-41 of Agrawal), until the current node includes more than a predetermined threshold number of attributes that indicate membership in a same class (see col. 4, lines 3-4 of Agrawal, as a number of values which is equal to or more than the threshold, each value of A from the set S is added).

As per claim 6, Agrawal teaches, wherein said partitioning step comprises the step of: for a current leaf node from among the leaf nodes of the decision tree (see col. 3, lines 44-47 of

Agrawal, as node's records which are split according to the best split test into partitions of records to form child nodes of the examined node, which also become new leaf nodes of the tree),

computing a lowest value of a gini index achieved by univariate-based partitions on each of a plurality of attribute lists included in the current leaf node (see col. 7, lines 12-16 of Agrawal, as to find the best split point for a node, the node's attribute lists are scanned to evaluate the splits for the attributes, the attribute containing the split point with the lowest value for the gini index is used for splitting the node's records).

As per claim 9, Agrawal and Ramaswamy substantially disclosed the invention as claimed. In particular, Agrawal teaches, wherein said partitioning step further comprises the steps of creating new child nodes for each of the two sets of ordered attribute lists (see col. 4, lines 7-11 of Agrawal, as a means for dividing the attribute list for B into new attribute lists corresponding respectively to the child nodes of the examined node). Agrawal does not explicitly disclose the steps of detecting subspace clusters of the records of the target class associated with the current leaf node; computing the lowest value of the gini index achieved by distance-based partitions on each of the plurality of attribute lists included in the current leaf node, the distance-based partitions being based on distances to the detected subspace clusters; and partitioning pre-sorted attribute lists included in the current node into two sets of ordered attribute lists based upon a greater one of the lowest value of the gini index achieved by univariate partitions and the lowest value of the gini index achieved by distance-based partitions. However, Ramaswamy discloses a system that teaches the claimed features "detecting subspace

clusters of the records of the target class associated with the current leaf node" as generating a set of clusters with generally uniform sizes and that fit in M, and wherein each cluster treats as a separate partition (see Ramaswamy page 432, col. 2, lines 25-27); "computing the lowest value of the gini index achieved by distance-based partitions on each of the plurality of attribute lists included in the current leaf node, the distance-based partitions being based on distances to the detected subspace clusters; and partitioning pre-sorted attribute lists included in the current node into two sets of ordered attribute lists based upon a greater one of the lowest value of the gini index achieved by univariate partitions and the lowest value of the gini index achieved by distance-based partitions" as a partition-based outlier detection algorithm that first partitions the input points using a clustering algorithm and computes lower and upper bounds on D^k "distance" for points in each partition; and we can use any of the L_p metrics like the L_1 or L_2 "euclidean" metrics for measuring the distance between a pair of points; and the square of the euclidean distance as the distance metric since it involves fewer and less expensive computations (see Ramaswamy page 428, col. 2, lines 7-10 and page 429, col. 2, lines 11-13; 36-46), which is similar to the description provided by the applicant (specification on page 33, lines 10-22). Applicant should duly note that Ramaswamy uses the distance of the k^{th} nearest neighbor of a point for a k and point p in which D^k (p) the distance of the k^{th} nearest neighbor of p, (see page 428, col. 1, lines 29-38). Therefore, it would have been obvious to a person of ordinary skill in the art at the time of the invention was made to modify the teachings of Agrawal with Ramaswamy so as to enable the detecting computing partitions and creating steps therein. Such a modification would allow the teachings of Agrawal and Ramaswamy the enhanced capability

to improve the efficiency of the method for building space splitting decision tree, therefore providing a quicker computation time of identifying small patterns for a given data analysis.

As per claims 16 and 17, Agrawal and Ramaswamy are discussed above. However, Agrawal does not explicitly disclose the claimed "wherein said computing step computes different Euclidean distance functions for at least some of the leaf nodes; wherein said computing step computes different Euclidean distance functions for each of the leaf nodes." Ramaswamy, on the other hand, discloses the claimed as we can use any of the L_p metrics like the L_1 or L_2 "euclidean" metrics for measuring the distance between a pair of points; and using the square of the euclidean distance as the distance metric since it involves fewer and less expensive computation; (see Ramaswamy page 429, col. 2, lines 11-13; 36-46), which is similar to the description provided by the applicant (specification on page 33, lines 10-22). Applicant should duly note, page 428, col. 1, lines 50-52, Ramaswamy teaches a distance of the k^{th} neighbor of the n^{th} outlier defines the neighborhood distance d. Therefore, it would have been obvious to a person of ordinary skill in the art at the time of the invention was made to modify the teachings of Agrawal with Ramaswamy so as to enable the detecting computing partitions and creating steps therein. Such a modification would allow Agrawal and Ramaswamy the enhanced capability to improve the efficiency of the method for building space splitting decision tree, and provide performance of the partition based algorithm relatively unchanged, (page Ramaswamy 437, col. 1, lines 17-18), therefore providing a quicker computation time of identifying small patterns for a given data analysis.

As per claim 18, Agrawal teaches, wherein said computing step comprises the steps of: for a current leaf node from among the leaf nodes of the decision tree (see col. 2, lines 21-26 of Agrawal, as a decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of records from the same class, the tree generally has a root node, interior nodes, and multiple leaf nodes where each leaf node is associated with the records belonging to a record class),

identifying relevant attributes of the current leaf node (see col. 3, lines 54-56 of Agrawal, as for each attribute which each leaf node includes one or more variables, such as histograms representing the distribution of the records at that leaf node);

computing a weight for each of the relevant attributes (see col. 7, lines 10-16 of Agrawal, as to find the best split point for a node, wherein the node's attribute lists are scanned to evaluate the splits for the attributes);

computing a confidence of the current leaf node (see col. 6, lines 8-12 of Agrawal, as for initially sorting (confidence) the numeric attribute lists once and future attribute lists created from the original lists will not need to be sorted again during the evaluation of split tests at each leaf node);

computing a centroid of the records of a majority class in the current leaf node (see col. 7, lines 31-33 of Agrawal, as a means for splitting index for the splitting criterion (A< or = v) at the examined node is computed at block thirty five); and

computing a weight of each relevant dimension of the current leaf node (see col. 7, lines 10-16 of Agrawal, as to find the best split point for a node, wherein the node's attribute lists are

scanned to evaluate the splits for the attributes, the attribute containing the split point with the lowest value for the gini index is used for splitting the node's records).

As per claim 19, Agrawal teaches, wherein an attribute is relevant when any node on a path from the root node to the current leaf node one of appears in a univariate test that splits the current leaf node (see col. 7, lines 12-16 of Agrawal, as a means for finding the best split point for a node the node's attribute lists are scanned to evaluate the splits for the attributes, the attribute containing the split point with the lowest value for the gini index is used for splitting the node's records).

As per claim 20, Agrawal teaches, wherein an attribute is relevant when any node on a path from the root node to the current leaf node one of appears in a univariate test that splits the current leaf node (see col. 7, lines 12-16 of Agrawal, as to find the best split point for a node the node's attribute lists are scanned to evaluate the splits for the attributes and the attribute containing the split point with the lowest value for the gini index is used for splitting the node's records).

As per claim 21, Agrawal teaches, wherein said identifying step comprises the steps of: for a current leaf node from among the leaf nodes of the decision tree (see col. 3, lines 40-41 of Agrawal, as a decision tree which is created by repeatedly splitting the records at each examined node, starting with the root node). Agrawal does not explicitly disclose the steps of computing a maximum distance of the current leaf node between a centroid of the current leaf node and any

of the records that are associated with the current leaf node; computing a minimum distance of the current leaf node between the centroid of the current leaf node and any of the records that are associated with other leaf nodes; forming the nearest neighbor set of the current leaf node to consist of only the other leaf nodes that have a corresponding minimum distance that is less than the maximum distance of the current node; and pruning from the nearest neighbor set of the current leaf node any nodes therein having a minimal bounding rectangle that contains the minimal bounding rectangle of the current leaf node. Ramaswamy, on the other hand, discloses an analogous system that teaches the claimed features "computing a maximum distance of the current leaf node between a centroid of the current leaf node and any of the records that are associated with the current leaf node; computing a minimum distance of the current leaf node between the centroid of the current leaf node and any of the records that are associated with other leaf nodes; forming the nearest neighbor set of the current leaf node to consist of only the other leaf nodes that have a corresponding minimum distance that is less than the maximum distance of the current node; and pruning from the nearest neighbor set of the current leaf node any nodes therein having a minimal bounding rectangle that contains the minimal bounding rectangle of the current leaf node" as a means for denoting the maximum distance between point p and rectangle R by maxdist(p,R) that is no point in R is at a distance that exceeds maxdist(p,R), the maximum distance between R and S, denoted by maxdist(R,S) is defined, the distance can be calculated using the following two formulae: maxdist(R,S) = $\sum_i^\delta x_i^2$ (see Ramaswamy page 430, col. 1, lines 1-14); as a means for denoting the minimum distance between point p and rectangle R by mindist(p,R), every point in R is at a distance of at least mindist(p,R) from p, the following is from mindist(R,S) = $\sum_i^\delta x_i^2$ (see Ramaswamy pages 429-430, cols. 2-1, lines 40-3); as a means

for defining the minimum and maximum distance between two MBRs, let R and S be two MBRs defined by the endpoints of their major diagonal, we denote the minimum distance between R and S by mindist(R,S) = $\sum_i^\delta x_i^2$ and similarly the maximum distance between R and S by maxdist(R,S) = $\sum_i^\delta x_i^2$ (see Ramaswamy page 430, col. 1, lines 11-26); further, in page 429, column 2, lines 36-46, Ramaswamy teaches the square of the euclidean distance as the distance metric since it involves fewer and less expensive computations, which is similar to the description provided by the applicant (see specification on page 33, lines 10-22). Therefore, it would have been obvious to a person of ordinary skill in the art at the time of the invention was made to modify the teachings of Agrawal with Ramaswamy so as to enable the detecting computing partitions and creating steps therein. Such a modification would allow Agrawal and Ramaswamy the enhanced capability to improve the efficiency of the method for building space splitting decision tree, and provide performance of the partition based algorithm relatively unchanged, (page 437, col. 1, lines 17-18).

As per claim 23, in addition to the discussion in claim 1, Agrawal further teaches wherein said method is implemented by a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform said method steps (see col. 10, lines 29-32 of Agrawal, as a means for resulting program and having computer-readable code means may be embodied or provided within one or more computer-readable media, thereby making a computer program product, the computer readable media may be, for instance a fixed "hard" drive, diskette, optical disk, magnetic tape, semiconductor memory such as read-only memory "ROM").

As per claim 24, Agrawal teaches a method for building a decision tree from an input data set, the input data set comprising records and associated attributes, the attributes including a class label attribute for indicating whether a given record is a member of a target class or a non-target class, the input data set being biased in favor of the records of the non-target class, the decision tree comprising a plurality of nodes that include and leaf nodes (see col. 3, lines 40-48 of Agrawal), said method comprises the steps of constructing the decision tree from the input data set, based upon multivariate subspace splitting criteria (see col. 3, lines 40-44 of Agrawal, as the decision tree which is created by repeatedly splitting the records at each examined node starting with the root node at any examined node a split test is determined to best separate the records at that node by record class and using the attribute lists);

classifying and scoring the records, based upon the decision tree and the nearest neighbor set of nodes (see col. 2, lines 44-46 of Agrawal, as a decision tree which can be used to screen future applicants by classifying them into the high or low risk categories); and

identifying a nearest neighbor set of nodes for each of the leaf nodes based upon a respective closeness of the nearest neighbor set of nodes to a target record of the target class (see col. 6, lines 52-54 of Agrawal, as the records at each new leaf node which are checked at block twenty three to see if they are of the same class). Agrawal does not explicitly disclose the steps of respectively measured by distance functions computer for each of the leaf nodes; identifying a nearest neighbor set of nodes for each of the leaf nodes based upon a respective closeness of the nearest neighbor set of nodes to a target record of the target class, as respectively measured by distance functions computed for each of the leaf nodes. Ramaswamy, on the other hand,

discloses an analogous system that teaches the claimed features "respectively measured by distance functions computed for each of the leaf nodes" as a means for using the square of the euclidean distance as the distance metric since it involves fewer and less expensive computations (see Ramaswamy page 429, col. 2, lines 36-46), which is similar to the description provided by the applicant (specification on page 33, lines 10-20). Applicant should duly note, that Ramaswamy uses the distance of the $k^{th.}$ neighbor of the $n^{th.}$ outlier to define the neighborhood distance d (see Ramaswamy page 428, col. 1, lines 50-52). Therefore, it would have been obvious to a person of ordinary skill in the art at the time of the invention was made to modify the teachings of Agrawal with Ramaswamy so as to enable the detecting computing partitions and creating steps therein. Such a modification would allow the teachings of Agrawal and Ramaswamy the enhanced capability of improving the accuracy and the efficiency of the method for building space splitting decision tree, and provide performance of the partition based algorithm relatively unchanged, (see Ramaswamy page 437, col. 1, lines 17-18).

As per claim 25, in addition to the discussion in claim 24, Agrawal further teaches wherein said method is implemented by a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform said method steps (see col. 10, lines 29-32 of Agrawal, as a means for resulting program and having computer-readable code means may be embodied or provided within one or more computer-readable media, thereby making a computer program product, the computer readable media may be, for instance a fixed "hard" drive, diskette, optical disk, magnetic tape, semiconductor memory such as read-only memory "ROM").

Allowable Subject Matter

6. Claims 7, 8, 10-15 and 22 are objected to as being dependent upon a rejected base claim, but would be allowable if rewritten in independent form including all of the limitations of the base claim and any intervening claims.

The prior art of record does not teach or suggest in combination with other elements, wherein the gini index is equal to 1-(P_n).sup.2-(P_p).sup.2, P_n being a percentage of the records of the non-target class in the input data set and P_p being a percentage of the records of the target class in the input data set as recited in claim 7.

The prior art of record does not teach or suggest in combination with other elements, wherein the percentage of the records P_p in the input data set is equal to W_p*n_p/(W_p*n_p+n_n), W_p being a weight of the records of the target class in the input data set, n_p and n_n being a number of the records of the target class and a number of the records of the non-target class in the current leaf node, respectively as recited in claim 8.

The prior art of record does not teach or suggest in combination with other elements, wherein said detecting step comprises the steps of: computing a minimum support (minsup) of each of the subspace clusters that have a potential of providing a lower gini index than that provided by the univariate-based partitions;

identifying one-dimensional clusters of the records of the target class associated with the current leaf node;

beginning with the one-dimensional clusters, combining centroids of K-dimensional clusters to form candidate (K+1)-dimensional clusters;

identifying a number of the records of the target class that fall into each of the (K+1)-dimensional clusters; pruning any of the (K+1)-dimensional clusters that have a support lower than the minsup as recited in claim 10.

Claims 11-13 further limit the subject matter of claim 10.

The prior art of record does not teach or suggest in combination with other elements, wherein said step of computing the lowest value of the gini index achieved by distance-based partitions comprises the steps of:

identifying eligible subspace clusters from among the subspace clusters, an eligible subspace cluster having a set of clustered dimensions such that only less than all of the clustered dimensions in the set are capable of being included in another set of clustered dimensions of another subspace cluster;

selecting top-K clusters from among the eligible subspace clusters, the top-K clusters being ordered by a number of records therein;

for each of a current top-K cluster,

computing a centroid of the current top-K cluster and a weight on each dimension of the current top-K cluster; and

computing the gini index of the current top-K cluster, based on a weighted Euclidean distance to the centroid; and

recording a lowest gini index achieved by said step of computing the gini index of the current top-K cluster as recited in claim 14.

The prior art of record does not teach or suggest in combination with other elements, wherein each of the plurality of pre-sorted attribute lists comprises a plurality of entries, and said step of partitioning the pre-sorted attribute lists comprises the steps of:

determining whether univariate partitioning or distance-based partitioning has occurred;

creating a first hash table that maps record ids of any of the records that satisfy a condition A=v to a left child node and that maps the record ids of any of the records that do not satisfy the condition A=v to a right child node, A being an attribute and v denoting a splitting position, when the univariate partitioning has occurred; creating a second hash table that maps the record ids of any of the records that satisfy a condition Dist(d, p, w)=v to a left child node and that maps the record ids of any of the records that do not satisfy the condition Dist(d, p, w)=v to a right child node, when the distance-based partitioning has occurred, d being a record associated with a current subspace cluster, p being a centroid of the current subspace cluster, and w being a weight on dimensions of the current subspace cluster;

partitioning the pre-sorted attribute lists into the two sets of ordered attribute lists, based on information in a corresponding one of the first hash table or the second hash table;

appending each entry of the two sets of ordered attribute lists to one of the left child node or the right child node, based on the information in the corresponding one of the first hash table or the second hash table and information corresponding to the each entry, to maintain attribute

ordering in the two sets of ordered attribute lists that corresponds that in the pre-sorted attribute lists as recited in claim 15.

The prior art of record does not teach or suggest in combination with other elements, wherein said classifying and scoring step comprises the steps of: for each of the plurality of nodes of the decision tree, starting at the root node,

evaluating a Boolean condition and following at least one branch of the decision tree until a leaf node is reached;

classifying the reached leaf node based on a majority class of any of the predetermined attributes included therein;

for each node in the nearest neighbor set of nodes for the reached leaf node,

computing a distance between a record to be scored and a centroid of the reached leaf node, using a distance function computed for the reached leaf node; and

scoring the record using a maximum value of a score function, the score function defined as conf/dist(d,p,w,), wherein the conf is a confidence of the reached node, d is a particular record associated with a current subspace cluster, p is a centroid of the current subspace cluster, and w is a weight on dimensions of the subspace cluster as recited in claim 22.

Prior Art

7. The prior art of record and not relied on upon is considered pertinent to applicant's disclosure. Agrawal et al. U.S. Patent No. 6,360,277 relates information is classified using a directed acyclic graph. Kothuri et al. U.S. Patent No. 6,381,605 relates to a system and methods are provided for indexing multi-dimensional data.

Contact Information

8. Any inquiry concerning this communication from examiner should be directed to Jean Bolte Fleurantin at (703) 308-6718. The examiner can normally be reached on Monday through Friday from 7:30 A.M. to 6:00 P.M.

If any attempt to reach the examiner by telephone is unsuccessful, the examiner's supervisor, Mrs. KIM VU can be reached at **(703) 305-8449.** The FAX phone numbers for the Group 2100 Customer Service Center are: ***After Final*** **(703) 746-7238,** ***Official*** **(703) 746-7239,** and ***Non-Official (703) 746-7240***. NOTE: Documents transmitted by facsimile will be entered as official documents on the file wrapper unless clearly marked "***DRAFT***".

Any inquiry of a general nature or relating to the status of this application or proceeding should be directed to the Group 2100 Customer Service Center receptionist whose telephone numbers are **(703) 306-5631, (703) 306-5632, (703) 306-5633**.

Jean Bolte Fleurantin

September 22, 2003

JBF/